Lisboa School of Economics & Management

Masters in Applied Econometrics and Forecasting

Correcting For Attrition in Panel Data Using Inverse Probability Weighting:

An application To the EU15 Bank System

Lutcy Afonso

February 2015

Lisboa School of Economics & Management

Masters in Applied Econometrics and Forecasting

Correcting For Attrition in Panel Data Using Inverse Probability Weighting:

An application To the EU15 Bank System

Lutcy Afonso

Advisor: Pierre Hoonhout

February 2015

Abstract:

This thesis discusses techniques to correct for the potentially biasing effects of missing data. We apply the techniques on an economic model that explains the Net Interest margin (NIM) of banks, using data from 15 countries that are part of the European Union (EU15) banking system. The variables that describe banks cover the period 2004 and 2010. We use the variables that were also used in Valverde and Fernández (2007). In addition, also macroeconomic variables are used as regressors. The selection that occurs as a consquence of missing values in these regressor variables is dealt with by means of Inverse Probablity Weighting (IPW) techniques. The weights are applied to a GMM estimator for a dynamic panel data model that would have been consistent in the absence of missing data.

Keywords: General Method of Moments; Inverse Probability weighting; Net interest margin; Sample selection.

Resumo:

Esta dissertação analiza técnicas de correção do efeito do enviesamento que pode ocorrer no caso dos dados utilizados apresentarem valores em falta. Tais técnicas serão aplicadas a um modelo económico para caracterização da margem líquida de juros (MLJ) bancária, utilizando dados provinientes 15 países que pertencem ao sistema bancário da União Europeia (UE15).

As variáveis que caracterizam os bancos são observados entre de 2004 e 2010. E são escolhidas seguindo Valverde and Fernández (2007). Adicionalmente aos regressores são acrescentadas algumas variáveis macroeconómicas. A seleção proviniente da falta de alguns valores para os regressores é tratada através da ponderação probabilistica inversa. Os ponderadores são aplicados a estimadores GMM para um modelo de dados de painel dinámico.

Acknowledgments:

I would like to express my gratitude to Pierre Hoonhout for his guidance throughout this study. I am also grateful to Luís Santos for the very useful tips he gave me about Stata. I also would like to thank Dra. Maria Hirondina Maciel da Silveira Duarte, Ana Margarida Martins Amaral and Rita de Cássia Botelho Pereira.

A special gratitude to my family, Ana Cristina F.R.F., and to my closest friends for their comprehension and motivation.

Contents

List of Tables

CHAPTER 1

Introduction

In recent years we have witnessed the wide impact that the financial system has on the economy. The main function of the financial system is to facilitate the allocation and use of economic resources. The 2007 crisis highlighted that the financial system also has a negative impact on the economy. In financial theory, there are two perpectives of the financial system. We focus on the global financial system, with an emphasis on the relationship between European countries. This group of countries is strongly linked by financial markets (e.g. stock exchange which includes all financial[1] institutions, borrowers and lenders in the world economy). What matters to us most is financial system stability. A stable system induces a better allocation of resources, mobilizes savings, reduces risks and facilitates transactions. Within this financial system we have the banks. These are the channel of the transations of resources. If this channel becomes turbulent, this has destabilizing effects. This can have several consequences: the reduction of resources within the banks, reduction of the interest rate and also a negative shock in the stock market. Under these circumstances, we cannot expect the banking system to positively influence the economy. As we have witnessed, it can have severe negative effects on the performance of the real economy, by destabilizing other non-financial institutions. A key role is played by the banks that have been considered "too big to fail". It is particularly this group of banks that make for an interesting analysis, especially with respect to management behavior, such as the impact of excessive risk in its performance.

In this thesis, we study the determinants of the so-called Net Interest margin (NIM) of banks, using data from 15 countries that are part of the European Union (EU15) banking system. The variables that describe banks cover the period 2004 and 2010. We use the variables that were also used in Valverde and Fernández (2007). In addition, also macroeconomic variables are used as regressors. The final economic model can

[1]For more detail of concept, types and definition see: http://www.investopedia.com/walkthrough/corporate-finance/1/financial-institutions.aspx

be described as a dynamic panel data model. In this model, we deal with the problem of missing data in some of the regressors.

The remainder of the thesis is organized as follows. Chapter (2) gives a brief overview of the relevant literature in the context of the economic model that we use. The data set is described in Chapter (3). The natural start of our analysis is a static linear regression model, which is described in Chapter (4). The extension to static and dynamic panel data models is described in Chapter (5). In chapter (6), we apply the inverse probability weighting method to the dynamic panel data model. Finally, chapter (7) concludes.

CHAPTER 2

The Economic Model

Recent history has shown that it is important to analyse the experience of financial intermediares (e.g. the bank system) under financial turbulance. The experience of banks can be summarized by their financial performance. This leads us to study what determines bank performance in the context of an economic crisis.

We measure the financial performance of banks by means of their Net Interest Margins (NIM). Facing a crisis, the banks will need to increase those, leading them to be more sensitive to interest rate fluctuations. In the presence of a crisis, these interest rates may grow faster than economic growth decreases. This reflects the contagion effect from the financial system to banks that can subsequently reduce the pace of economic growth.

These channels of contagion can both be measured by the spread. The first emphasizes the intermediation role of the financial system in providing credit to the real sector and the second gives special importance to the transition of shocks and disruptions of financial system to the real economy.

In the period of analysis (more precisely, after 2007) , the response of the bank system to the tight market credit conditions (that resulted from the crisis) was relatively similar for each country. The implementation of several policies, as mergers, acquisitions and takeover bids (even thought, with this syncronization but without a strong regulation, as was the case in this period, seeKasman et al. (2010), one cannot conclude that the policy results were the same, even if the policies were similar in all countries This is not a good signal to the european union as a whole and its political integration process.

The macroeconomic theoretical and empirical literature both suggest that well-designed and efficiently managed budget institutions can play a central role in the success of fiscal consolidations. We can enlarge the theory to country-bank institutions that can support the financial consolidation process. This suggests that we can use the banks systems to identify the economic performance and synchronization of EU15.

In Ho and Saunders (1981), the bank is viewed as a risk averse dealer between lending and deposit rate, and is specialized only in intermediation activities (the traditional actvities of a bank). That theory developed a model of bank margins where a pure spread estimation is given by the constant term of a seminal static model [1]. This pure spread is also called net interest margin(NIM)[2].

When the spread is negative the bank loses money and it was shown in the NIM[3] model that the pure spread is explained by four instable factors: managerial risk aversion; the size of transactions undertaken by the banks; bank market structure and the volatility of interest rates. This results in business uncertainty faced by the banks. Allen (1988) improvedHo and Saunders (1981) model by including different types of credits and deposit (portfolio) to managing risk exposure. Note that, for simplicity, only two interest rates were considered: one for traditional and one for non-traditional activities[4].

Saunders and Schumacher (2000)investigate the impact of the structure of bank competition and interest rate volatility on the net interest margins. In this model, the pure margin is also explained by the degree of bank concentration, which can vary from country to country. In that analysis, they also suggest that policy can respond to the variability of the interest rate, by using government's macroeconomic policies such as changes of the reference interest rate to manage the cost of intermediation, or consequently a positive effect reducing banks margins. The problem of this intervention is that nowadays only the ECB can make the decision to change the interest rate and the dataset we use also contains earlier data (when EU15 countries had monetary autonomy).

Demirgüç-Kunt et al. (2003) measured the impact of regulations, market structure and national institutions on the cost of intermediation by net interest margin and bank

[1]The estimates of pure spread($\hat{\delta_0}$) are taken from $M_i = \delta_0 + \delta_1 IR_i + \delta_2 OR_i + \delta_3 DP_i + U_i$.

[2]*Ho and Saunders (1981) definition: The interest margin is defined here as the spread between the interest revenue on bank assets and interest expense on bank liabilities as a proportion of average bank assets–this margin is also called the banker's mark-up*

[3]To make clear, it is defined as the difference between the interest paid by the creditor and the interest received by the savers on their deposits.

[4]These two prices comes to facilitate the analysis of ceteris paribus, when the price (interest rate or spread) of traditional activity decreases, ceteris paribus, it is expected that the price of non-traditional activity increases.

overhead expenditures. Beck et al. (2006) studied the relationship between national bank concentration, competition and crisis. Their results show that bank concentration tends to reduce the probability of experiencing an economic crisis and boosts bank profits. Maudos and de Guevara (2004) expanded the Ho and Saunders (1981) model by taking into account a bank's operating cost explicitly (they consider the crucial elements affecting this margin). They use direct measurements of the degree of competition in the different countries (concentration index and market power). The analysis concludes that countries with a less concentrated and more competitive banking system should display less persistency in NIM.

Valverde and Fernández (2007) applied the seminal model of Ho and Saunders (1981) to a dynamic multi-output framework. This simply means they vary (for each equation explains a different dependent variable for the same regressors or the explanatory variables is equal in all equation, bank-specific variables) dependent variable (addition to the spread, consider four dependent variables) to identify different types of of interest margin [5]. Claeys and Vennet (2008) They come into play either when we have an economic theory to test or when we have a guessing relationship that has some importance for banks decisions or policy analysis, introducing the macroeconomic environment and the influence of state-owned[6] banks. amongs others conclusion, in the both cases, they conclude that the presence of macroeconomic enviroment and state-owned banks can explain significantly in the model of net interest margin. We focus on analysis in the net interest margin (NIM), which is a good proxy in the examination of bank performance.

[5]To give some idea, the model looks some type of dynamic Seemingly Unrelated regression, where besides Nim as dependent variable, they include four more definition of bank margin, keeping vector x's constant in all equation.

[6]A financial institution that has been chartered by a state to provide commercial banking. An exemple in the case of Portugal state-owed bank is Caixa Geral Depositos, because it is financed by public money.

CHAPTER 3

The Data

In this Chapter we describe the data used for our research. It proved difficult to obtain a sufficient amount of Banking data to be representative. The data on the bank level were taken from the Bankscope database and macroeconomic data from the International Monetary Fund (see variables in the Appendix) Table (A1). The bank information for the European Union of fifteen (EU15) covered the years 2004 to 2010.

For the bank data, there is a problem of information reporting, related to dimension: bigger banks provide the information we need, once they are in the stock market, which is not the case for most of the smaller banks. If we try to add all the banks recorded on the database and use them to build a sample, there might be a selection bias to account for. For that reason, we exclude smaller banks that mostly do not report any information. To assess the relevance of banks for inclusion into our sample, we calculate the level of concentration of the banking system for each country and count the number of banks that are part of this level. In some cases this allows us to include small banks that are listed on stock exchanges.

Overall, we have in total $N = 267$ banks in our final panel dataset. These banks can be considered "too big to fail", in the sense that the biggest banks in each EU15 country are included into the sample. We do not report the values for the level of concentration (Gini Index) because it was used only to set up the number of banks by country. The bank specializations are as follows: commercial, cooperative, investment, real estate,and savings.

In the analyzed time period, the sample contains banks that are not observed in all time periods, for example, banks that failed (bankruptcy). This introduces the missing data in the sample. The dependent variable, however, is observed throughout. The restriction of not becoming bankrupt (banks-alive) will be useful in Chapter (6).

The complete table for the summary statistics of the raw of variables used in the analysis can be found in Table A1(Appendix) and summary statistics Table A2(Appendix). For each country in our panel survey, the number of observation is given in the Table

A3 (Appendix). In the section of Bank-specific variables, the first variable is the dependent variable labeled as NIM. The regressors that are measured at the bank-level include the determinants of spread. The macroeconomic variables include Inflation and GDP growth.

The bank-specific variables include the Cost to Income ratio (Cost), the Loans to Total Assets ratio (Loan), the Loan Loss Provisions to Gross Loan ratio (Lloss), Equity to Total Assets ratio (Etar), the Total Earning Assets to Total Assets ratio (Teatr) and Off-balance items (Oba). These variables proxy management quality/efficiency. To measure risk exposure: the Credit risk (Crisk), the asset risk: the liquidity risk (Liqrisk); and the difference between the interbank market rate and the interest rate for customer deposits: the Interest rate risk (Irrisk). At times, we also include country-dummy variables.

Not all variables in our sample are observed in all time period of analysis, ($t = 1, \ldots, T$). The variables exhibit patterns of missing data that are different for each country, as shown in the Table A2 (Appendix)

CHAPTER 4

The Static Linear Regression Model

The initial empirical analysis described in this Chapter uses time series data to analyse the theory about the NIM: we estimate the relationship between the independent variable NIM, using bank-specific risk variables and macrocecomics variables as regressors. The linear regression model is an obvious first step in this analysis. After estimation, we can then test certain aspects of the economic theory, or test the effects of a certain monetary policy. For example, a test that assesses the effect of inflation on NIM. In all cases, the regressors will have to be considered random variables in this economic context.

The sample means are displayed in Table(4.0.1). A more informative analysis is obtained by studying their contemporaneous relationships.

In general, the static regression model elucidates the effect of a change in x_t at time t on y_t (it is called static model because all explanatory variables are dated contemporaneously and only with subscript t). Obviously, the analysis proceeds under the ceteris paribus condition, because we have several explanatory variables in the static regression model. The basic linear regression model allows us to answer a wide range of questions. For convenience we omit the i subscript to be used in the next Chapter, provided that its omission does not cause any confusion. In period t, our linear model can be described as it follows:

$$y_t = \beta_0 + x_t\beta_1 + z_t\beta_2 + u_t,\ t = 1, \ldots, T. \tag{1}$$

where y_t is the determinant net interest margin, $\beta_0 = \delta_0$ is the pure spread at instant t, the vector x_t includes all contemporaneous bank-specific regressors. The vector z_t is includes all contemporaneous macroeconomic regressors that describe the country-specific macroeconomic situation at that time period. u_t is disturbance error term.

The parameters $\beta_1$and β_2 correspond to the parameters of interest. Whether these parameters can be estimated consistently depends on the time series assumptions made. A crucial assumption is stated in terms of conditional expectations: it requires that u_t is be uncorrelated with both $(x_{sj}, z_{sj}) \equiv B_{sj}$ in each time period. That is,

$$E(u_t|B_{t1}, \ldots, B_{tk}) = E(u_t|B_t) = 0,\ t = 1, ..., T, \tag{2}$$

The above condition is sufficient to assure consistency[1] of the OLS estimator. Recall that in Eq.(1), for period t, we can estimate the linear vector of β_1 and β_2 to explain the effect of the bank-specific regressors and the macroeconomic regressors on NIM. We can also measure the individual country effect by including country dummy variables. All explanatory variables are available as percentages except for the dummy variables. We can estimate the single-equation linear regression model in Eq.(1) under assumption of Strict Exogeneity, i.e. Eq.(2), by ordinary least squares (*Pooled*[2] *OLS*). We can obtain a different regression model estimate if we assume independent cross sections for each i and t. The results are given in Table(4.0.2).

Before concluding anything from these estimation results, it is important to remember that we are interpreting and testing hypotheses about the general case of population parameters. For example, for the regressor $loan_t$, we obtain: a 1% increase in loan leads to an increase of NIM of about 1.12%. Moreover, the effect of loan on NIM is statistically significant.

TABLE 4.0.1. EU15 average over time

year	nim	cost	loan	lloss	etar	teatr	oba	crisk	irrisk	liqrisk	gdpg	hipc
2004	1.780	60.495	57.689	4.479	6.058	84.039	31.570	1.907	143.025	47.658	0.043	0.018
2005	1.627	57.065	56.941	6.565	6.532	78.859	30.216	1.441	112.197	55.593	0.037	0.021
2006	1.609	56.766	57.658	3.605	6.402	78.536	43.887	1.083	119.265	50.619	0.052	0.021
2007	1.529	56.560	57.296	4.017	6.164	79.808	37.382	1.365	103.450	55.511	0.052	0.021
2008	1.547	69.277	58.216	2.598	5.256	79.702	76.706	1.156	97.172	52.914	-0.001	0.035
2009	1.542	58.292	58.112	5.081	5.590	79.973	95.640	1.185	95.314	44.801	-0.057	0.007
2010	1.426	60.979	58.058	9.456	5.552	79.914	25.277	1.155	90.282	42.203	0.039	0.019

[1]It is implicity that a model in the Eq.(1) no lags is need to explain the expected value of y_t, meaning $E(y_t|B_t, B_{t-1}, \ldots, B_1) = B_t\delta$.

[2]The brief view of Pooled here, is in a concept of a panel data set that follow the same group of banks over time. In that context, the consistency of estimator depend on assumption Eq(2), rank condition, $rank[\sum_{t=1}^{T} E(B_t^{'}B_t)] = K$ and no serial correlation $E(u_t u_s B_t^{'} B_s) = 0$, fot $t \neq s$; $t, s = 1, \ldots, T$.

TABLE 4.0.2. Pooled OLS

Dependent variable NIM	
Variables	Coefficient
cost	-0.00033
loan	0.01122***
lloss	0.00149
etar	0.06502***
teatr	0.00972**
oba	0.00001
crisk	0.01874***
irrisk	0.00064**
liqrisk	-0.00349***
gdpg	-0.39917
hipc	-1.43544
Constant	0.16998
Observations	830
R-squared	0.333
Robust standard errors	
*** p<0.01, ** p<0.05, * p<0.1	

CHAPTER 5

The Linear Panel Data Model

Panel data analysis represents a hybrid of cross-sectional analysis and time series analysis. With time series analysis, we identify one or more banks and observe them over time and allow us to study the dynamics. On the other hand, the presence of heterogeneity (in cross-sectional units) that is unobserved (and hence part of the error term) induces endogeneity of the regressors: the possibility of correlation between covariates and the errors. This is a violation of the exogeneity assumption, and leads OLS to be inconsistent.

As emphasized by many authors, like Kasman et al. (2010), and others, many of the fundamental issues that dominated earlier work based on bank market data remains important, especially issues concerning the conditions necessary for identifiability of causal economic relations, regulation and risk of capital with NIM.

We can start with the single linear economic model proposed in Eq(1). As a baseline analysis that improves on the linear regression model discussed above, we consider a static linear panel data model. The regressors now form the vector x_{it}, a vector of variables that take different values both over time and over banks. No lagged dependent variables are included in the regressors. This model already provides an improvement of a model proposed by Valverde and Fernández (2007). The model reads:

$$(3) \qquad \begin{cases} y_{it} = x_{it}'\beta_1 + z_t'\beta_2 + u_{it} \\ u_{it} \equiv \mu_i + v_{it};\ i = 1, ..., N;\ t = 1, ..., T \end{cases},$$

where x_{it} is $K \times 1$ vector of bank-specific time-varying variables, z_t is $K \times 1$ vector of macroeconomics time-varying variables, $\beta_i,\ i = 1, 2$ a $K \times 1$ vector of parameters, μ_i denotes the unobserved bank effect (or country-specific effect), an unobservable that doesn't change over time but is different for different banks. The macroeconomic variables included in model are time-varying only. The two[1] components of the error term,

[1]The macroeconomic variables capture time effect, is double count if we include lambda in the estimation, for convenience we excluded.

u_{it}, are assumed to be uncorrelated with each other. The variables in x_{it} are possibly endogenous, as they may be correlated with μ_i and v_{it}. The macroeconomic variables in z_t are assume strictly exogenous and possible correlated with time invariant specific effect. To simplify notation, let $\delta \equiv (\beta_1, \beta_2)$ be the vector of parameters of interest and $B_{it}^{'} \equiv (x_{it}, z_t)^{'}$. The error component model structure across the system of equations then becomes $y_{it} \equiv B_{it}^{'}\delta + u_{it},\ u_{it} = \mu_i + v_{it},\ i = 1, \ldots, N;\ t = 1, \ldots, T$. The regressors x_{it} could include lagged dependent variables $y_{i,t-1}$.

5.1. Testing For Endogeneity

Most of economics variables are related with error term because they are simultaneously determined, amongst other sources of endogeneity. Of course endogeneity leads to inconsistency of the OLS estimator of the parameter in Eq.(1). the estimates of β when estimated by ordinary least squares (OLS) do not converge in probability to the true parameter values.

Since we have at our disposal a panel data set, we can do better than OLS. For the moment, we only consider observations that are fully observed. Given that, in our model we expect that the error term contains an unobserved bank-specific effect that may well be correlated with the regressors included in the fixed effect specification in Eq.(3). It is important to keep in mind that Eq.(1) estimated by OLS or two stages least square $(2SLS)$ must generate the same error term in the absence of endogeneity.

Bearing in mind the source of the endogeneity, the traditional methodology to test correlation between x_{is} and u_{it} for any s and t as was suggested by Hausman (1978). Simply put, he proposed comparing two estimators of the parameters. One efficient and one consistent under the null-hypothesis. Combining those estimates, he derived a test of orthogonally between covariates and the error, $E(u_{it}|x_{it}) = 0$.

Under the null hypothesis, there exists a consistent estimator, the so-called Random Effects estimator (RE). This estimator is asymptotically normal and efficient. Under the alternative we have only a consistent estimator, the Fixed Effect estimator (FE). This estimator is not efficient under the null hypothesis. When we consider the properties of each estimator under the alternative hypothesis of $E(u_{it}|x_{it}) \neq 0$, the RE estimator will be biased and inconsistent and FE still consistent. Then, the general idea of the test is to contrast the two estimator, RE against FE.

In the literature, some authors have simply assumed that the error component is uncorrelated with vector of variables $E(\mu_i|x_{it}) = 0$. They run Generalized Least Square (GLS or RE)[2] which is consistent and asymptotically efficient under null hypothesis. They ignore the possiblity that the estimator β_{GLS} is inconsistent. To give some examples that investigate the NIM, consider Demirgüç-Kunt et al. (2003), Claeys and Vennet (2008) and Kasman et al. (2010). They all applied GLS without testing for the validity of the assumption assumed to obtain the consistent estimates.

An possible improvement on the static linear panel data model is a dynamic linear panel data model. Provided that our goal is to estimate a dynamic model, the standard Hausman test has some drawbacks. Firstly, FE is widely known to be biased and inconsistent for the dynamic model as the number of banks increase to infinity if the number of time periods (T) is kept fixed. In fact, that the inconsistency is of order $O(1/T)$, which means in our case, having a micro panel, that the size of bias cannot be ignored. That case was shown by Nickel (1981) and more recently extended by Phillips and Sul (2007) to incorporate processes with a unit root.

Secondly, if dynamics are important, both FE and RE estimators[3] will be inconsistent because of the failure of the Strict Exogeneity assumption. Note that, the strong assumption of exogeneity implies that there is no correlation between x_{is} and u_{it} for any s and t causes both FE and RE to be consistent, otherwise, their probability limits will differ.

The third reason is potentially even more serious: the RE estimator is usually implemented assuming constant variance assumption $(RE.3)$ under the null. It is for this reason that the RE estimator is more efficient than the FE estimator. But if this assumption is violated, the sum of squared residual that form of the F statistic is not valid. Moreover, it causes the Hausman test to have a nonstandard limiting distribution and the size of test can be undersized or oversized. For this reason, we will use a more robust Hausman-type test, to be discussed below.

As mentioned above, the choice between a FE and RE approach in the static linear panel data context depends on whether or not μ_i and x_{it} are correlated. See more detail of the derivation in Hausman (1978). He suggested a test based on the difference

[2]see Wooldridge ((2010, Ch.10)) under which condition needed for consistency.

[3]The valid to implement FE and RE estimator, it necessary to assume assumption of FE.1-FE.3 and RE.1-RE.3 for consistency, see Wooldridge (2010).

between the RE and FE estimates, which can be obtained after a few lines of algebra. We can compute the Hausman test-staristic as follows:

$$y_{it} = \beta x_{it} + u_{it},\ u_{it} = \mu_i + v_{it},\ t = 1, \ldots, T \tag{4}$$

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' [\widehat{Avar(\widehat{\beta}_{FE}\,)} - \widehat{Avar(\widehat{\beta}_{RE}\,)}]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \tag{5}$$

The statistic has an (asymptoticaly) chi-square (χ^2_k, *hereafter*) distribution, where k denotes the dimension of slopes of linear vector $\hat{\beta} = (\hat{\beta}_{FE},\ \hat{\beta}_{RE})$ and $\hat{h} = (\hat{\beta}_{FE} - \hat{\beta}_{RE})$.

The resulting Hausman test statistic based on the reduced[4] Eq.(4) where the vector x_{it} is banks specific variables observed during 2004 to 2010 yields observed χ^2_k statistic of 81.87 with $p - value = 0.000$. These do reject the null hypothesis that RE yield consistent estimators, or there is correlation between the regressors,x_{it}, and banks-specific effect[5](μ_i). See table below for the Hausman test. We will see below how this Hausman test can be generalized to lead to reliable results even if there is correlation between the regressors and the idiosyncratic part of the error term.

TABLE 5.1.1. Standard Hauman Test

Variables	(b) FE	(B) RE	(b-B) Difference	S.E.
cost	-0.00211	-0.00152	-0.00059	0.00017
loan	-0.00200	-0.00393	0.00193	0.00161
lloss	0.00397	0.00936	-0.00538	0.00242
etar	0.01975	0.03906	-0.01932	0.00257
teatr	-0.00882	0.00015	-0.00897	0.00858
oba	0.00002	0.00001	0.00001	0.00001
crisk	0.00091	0.01101	-0.01009	0.00995
irrisk	0.00015	0.00034	-0.00019	0.00007
liqrisk	-0.00227	-0.00282	0.00055	0.00038
chi2(9)=81.87 with Prob>chi2=0.000				
Nº Observation = 830				

4The test should be performed in context of strong within variation. In contrast, there is potential invalid test under small variation in within transformation, see Hahn et al. (2011). They suggest a re-sampling procedure that approximates the distribution H. Ours variables in analysis have not revealed that problem.

5In the Chapter(4) we put μ_i into the error term and that can cause serious problems of consistency.

5.2. The Fixed Effect Model

As the null-hypothesis of no correlation between the fixed effect and the regressors has been rejected, in this Chapter we analyse our economic model with the fixed effect (FE) estimator. As in Ho and Saunders (1981), the goal is estimating the pure spread. Using Eq.(3) it is necessary to impose the condition that country bank heterogenity is "constant" over time (μ). However, the inclusion of time specific effects turns out to lead to a near-multicollinearity problem in our application. It so happens that inclusion of time dummies implies that the estimates become very imprecise, especially on the macro-economic variables. This is not surprising, as these macro-economic variables are already included to pick up the same effect as the time dummies. For this reason, we exclude the time-dummies from the analysis below. That is, we only control for unobserved time-invariant heterogeneity ($\underline{\mu}$) Although the bank-efects μ_i are included in our model, we do not estimate them as we are only interested in β. To do so, we can use either the fixed effect estimator (FE) or the First difference estimator (FD). Both estimators eliminate the bank specific effect by a straightforward transformation. Which estimator should we use?

In our case we have time periods $T > 2$ and sufficient large, fixed effects estimation and first differencing produce different estimates and inference. Therefore, we need to choose our estimator on the basis of which of the two estimators is more likely to satisfy the assumptions that are behind their consistency. In particular, FD is consistent if u_{it} has a unit root, whereas FE is not. If there is no unit root, FE is more efficient. As we have no reason to suspect unit roots in our application, we to use the FE estimator.

The error composition is $u_{it} = \mu_i + v_{it}$. The structure of the variance of this compound error is given by the $cov(u_{it}u_{js}) = E(u_t u_j^{'}) = E[(\mu_i + v_{it})'(\mu_i + v_{it})]$. As panel data are available, we can estimate Eq.(3) by FE. Consistency of FE will depend on the Strict Exogeneity condition as in $Pooled - OLS$, i.e., $E(u_{it}|x_{it}) = E[(\mu_i + v_{it})|x_{it}] = 0,\ t = 1, ..., T$. This assumption is by no means trivial. If at least, banks heterogeneity is correlated with x_{it}, and/or v_{it}, the conditional expectation is unequal to zero, introducing biased in the FE estimator. Ignoring this for the moment, the FE estimation results are given in Table(5.2.1). These results can be interpreted in the usual way. For example, compared with the results of $Pooled - OLS$, the effect of variable $loan_t$ is less pronounced and insignificant. Conversely, the effect of spread is stronger and now significant.

TABLE 5.2.1. FE Model

Dependent variable NIM	
Variables	Coefficient
cost	-0.00204**
loan	0.00408
lloss	-0.00218**
etar	0.01953
teatr	-0.00892
oba	0.00002***
crisk	0.00110
irrisk	0.00015
liqrisk	-0.00223*
gdpg	0.04408
hipc	-1.35760
Constant	2.13631***
Observations	830
Number of nr	198
R-squared	0.039
Robust standard errors	
*** p<0.01, ** p<0.05, * p<0.1	

5.3. Dealing With Additional Endogeneity

Until now, all the estimation methods we assumed that the explanatory variables were strictly exogenous. There could still be endogeneity due to correlation of regressors with v_{it}. Dealing with that requires instrumental variables.

5.3.1 The Baltagi-Li Hausman Test. This Chapter develops a slightly different approach to the basic Hausman test. The problem at hand is the following: if we stop assuming that the regressors and v_{it} are uncorrelated, the standard Hausman test becomes invalid. We need an alternative test. In this case, we need to add to the FE-approach some additional endogeneity-solution that takes the form of instrumental variables (IV). The FE transformation wipes out the time constant error μ_i, but not the time-varying error v_{it}. Hence, this second source of endogeneity is not dealt with by the FE estimator. This motives us to consider a simple to test for endogeneity in a regression estimated via instrumental variables, where we add the to the vector z_t (see Table(A1), country specific variables) a set of credible instruments for the estimation of Eq.(4).

Inspection of the table of available variables reveals two problems: Firstly, we need valid instruments , and secondly, we need enough of them. See Table(A1). It is well-known that if $K < L$, in the notaion of Hayashi Hayashi (2000, p.202), the model is not identified. For this reason, we add more macroeconimc variables, and define the vector $Z_t \equiv (z_t, z_{t-1}, \triangle z_t)$. With this definition, we have a vector that consists of exogenous variables with dimension $1 \times K$ satisfying the moment conditions, $E(Z_t u_{it}) = 0$. Now, the usual matrix W of valid instruments is equal to Z, the condition in the estimation, $iv(Z_t)$. This implies that may utilize the $K = L$ moments condition and end up with an exactly identified model. With this solution of instruments, we cannot test the specification of our model using the overidentification test of Hansen, since that test requires $K > L$. Even so, whats matter is the quality of instruments. If $cov(x_{it}, z_t)$ is unequal to zero, we can use z_t as instruments of x_{it}.

As done by the other authors mentioned above, in the empirical literature on NIM it is customary to include only two macroeconomic control variables, (GDP) and ($HIPC$). Here we include more variables as instruments, and the main question is if the variables in Z_t are plausibly valid instruments, even disregarding the possibility that they are valid but weak (John Bound and Baker (1995)). Weak instruments can however be discoverd by reporting the $F - statistic$ in first stage regression (FF). Under null hypothesis, the vector of instruments $\beta_2 = 0$ in Eq.(3), a hypothesis that we would prefer to reject (i.e. the instruments are valid). Recall that the equation of interest is Eq.(4) and here the estimation of $\beta_2 = 0$ involves all macroeconomic variables in the Table(A1) as instruments only. This test is performed in the first stage of $FE - 2SLS$, and if we do not reject null hypothesis, it means that the generalized least squares estimator is asymptotically biased in direction of ordinary least squares, as discussed in Nelson and Startz (1990b). The latter paper also proposes the reporting of the partial $\bar{R}^2$ in the first stage ($P\bar{R}^2$) regression to evaluate the sensibility of sample size, but the problem of poor instruments does not reduce with enlargement of N (number of banks) under assumption of fixed instruments and normal error. More recently, Staiger and Stock (1997) proposed an alternative asymptotic approximation to the finite distribution of the $2SLS$ estimator[6] and their $t - statistic$ for a sample with 10 to 20 observations per

[6](In this second case, they propose the possibility to choose one of the estimators. In that case, their results support the LIML estimator, in many cases median unbiased.)

instruments. The asymptotic distribution can provide good approximation to sampling distributions, but this is not the only approach, and for some problems this is not the best approach, in the sense that it does not necessarily provide the most accurate approximation. We care about the small sample property, because in some case we could easily end up with a very small sample. Taking into account the sensitivity of the consistency of the FE estimator, it is necessary to consider under what conditions the RE estimator works better than the standard Hausman test.

As FE estimation can be generalized using instrumental variables, so can RE. To alleviate the problem of the flawed distribution of the Hausman test, we also consider an alternative approach to get a feasible estimates for the Error Component 2SLS (EC2SLS), to replace RE. The main result that we use here, are all available and demonstrated in Baltagi (2005). EC2SLS is to be preferred over RE with IV (G2SLS).

The above discussion implies that $E[(\mu_i)|x_{it}] = 0$, can be testing using a Hausman-like test where instead of using FE *vs* RE, we now use FE2SLS *vs* EC2SLS. In our implementation, we macroeconomic variables are as instruments. Instead of Eq.(4), Baltagi and Li (1992) suggest that the model and instruments used both are tranformed by variance. Then, the $EC2SLS$ estimator becomes estimatos of the equation, (and FE2SLS)

$$y^+_{it} = \beta_1 x^+_{it} + u^+_{it}, \tag{6}$$

with the instruments z^+_t, the vector of ($y^+_{it}, x^+_{it}, z^+_t$ *and* u^+_{it}) are obtained by premultiplying Eq.(4) by variance[7], see also Baltagi and Liu (2007) and Ahn (1996) alternatives[8] methods to obtain the Hausman test, wich showed that the asymptotic χ^2_k where k denotes the dimension of parameter vector β, then the testing can be obtained using an artifial equation. If the model is identified, estimating robust Eq.(6) by 2SLS, where under null hypothesis we obtain the same Hausman statistic as $\hat{h}_{2SLS} = (\hat{\beta}_{FE2SLS} - \hat{\beta}_{EC2SLS})$. That result is just appling $FE - 2SLS$ and Error Component in Eq.(4) repectively, $(\hat{\beta}_{FE2SLS}, \hat{\beta}_{EC2SLS})$ to obtain $\hat{h}_{2SLS}$ consistently. The chi-square with

[7]To be more especific, see Baltagi (2005, Eq.(7.16))

[8]I considerer these aproach, which is motivated on Mundlak (1978), but these alternative methods reveals poors results, especially very high variance. Although the result of GMM suggested by Ahn (1996) reject the null hyphotesis, but statistic which guarantee the consistency were rated poor values.

k degree freedom converges whether or not assumption $RE.3$ holds[9]. For an application[10], without going into proofs, Baltagi (2005) gives us both possibilities. Assuming that the results of Baltagi (2005) hold, the observed Wald test statistic equalst 25.64, which is larger than χ^2_k, with $k = 9$ at 5% level significance and p-value $p = 0.0023$, see Table(5.3.1).

With these two results, the conslusion must be that we cannot ignore the correlation between (μ_i, v_{it}) and x_{it}.

TABLE 5.3.1. Hausman Test (EC2SLS)

Variables	(b) FE-2SLS	(B) RE-EC2SLS	(b-B) Difference	S.E.
cost	-.00565	.00060	-.00626	.00455
loan	.01952	.03161	-.01208	.03173
lloss	-.01219	-.01102	-.00118	.01338
etar	.08103	.05973	.02131	.06894
teatr	.05397	.00217	.05180	.07798
oba	.00042	.00039	.00004	.00018
crisk	.014900	.03506	-.02016	.05036
irrisk	-.00102	.00169	-.00273	.00086
liqrisk	.00310	.00109	.00201	.00598
chi2(9)=25.64 Prob>chi2=0.0023				
Nº Observation = 830				

TABLE 5.3.2. FE-2SLS :First Stage statistic

statistics	cost	loan	lloss	etar	teatr	oba	crisk	irrisk	liqrisk
$P\bar{R}^2$	0.0397	0.0529	0.1266	0.0357	0.0367	0.0295	0.2294	0.0392	0.0481
$FF\ test$	1.70	2.30	5.96	1.52	1.57	1.25	12.24	1.68	2.08
$Prob > FF$	0.0470	0.0035	0.0000	0.0910	0.0778	0.2284	0.000	0.0505	0.0095

5.4. Dynamic Panel Data Model

In many applications it is beneficial to allow for dynamics between the economic variables. This means that the effect of one regressor variable on the dependent variables is allowed to be non-instantaneous. That is, the effect is allowed to take some time to be fully achieved. A static model does not allow for this possibility. This

[9]See also Wooldridge (2010)"when assumption RE.3 fail, we can run Pooled OLS in Eq.(10.88).

[10]See also empirical exemple in Baltagi (2005)Table 7.1 where he uses simple stata command in Eq.(4) to obtain $\hat{h}_{2SLS}$.

idea is implemented by including a lagged dependent variables as a regressor into the model.

In the context of panel data, the survey in generally contains sufficient information about earlier time periods to allow for the dynamicsrelationship of interest. If the relationship is dynamic in reality, a static panel data model will yield inconsistent estimates of the parameters of interest. Therefore, even if the coefficients on the lagged variables are not the main interest, introducting dynamics in the regression may be crutial for consistent estimates of the other parameters.

Using the notaion introduced earlier, we consider the following dynamic panel data model, keeping the FE structure of the error term,

$$y_{it} = \rho y_{i,t-1} + x'_{it}\beta_1 + z'_t\beta_2 + u_{it},\ i = 1, ..., N;\ t = 2, ..., T. \tag{7}$$

The inclusion of lagged variable puts in effect a concept of time series data of a stationary process, assuming[11] $|\rho| < 1$. The latter assumption avoids problems associated with unit roots. However, we still consider the possibility of endogenous regressors and their IV solution. Introducing dynamics does not solve the endogeneity problems that were present before.

By definition, y_{it} is a function of μ_i which is contained in the error term, and consequently $y_{i,t-1}$ is also a function of μ_i. This implies that the regressor is correlated with error term, which implies the need of IV methods. Since we can relax the assumption of Strict Exogeneity using GMM, motivated by the fact of this being asymptotically more efficient (as there exists a positive definite weighting[12] matrix which produces a smaller asymptotic variance) among all of the possibilities, and solving the problem of the instruments with the most distant observations, we consider this approach more credible than 2SLS procedure discussed above.

In addition to efficiency, $FE - 2SLS$ is a particular case of GMM, since in this Chapter we are concerned to address most general questions, we can also add that with respect to inconsistency of the estimator, the system-GMM reduce that problem. The reduction of the finite sample bias, may be it is better to use GMM estimator into

[11]Usually, this became a problem when we talk about pure time series analysis, or in the panel data case when $T \to \infty$ and N fixed.

[12]Matrix of the linear full set moment conditions, see Wooldridge (2010, p.213).

the equation to obtain parameter estimates. The most serious problem from the finite sample bias is illustrated in the empirical application shown by Hayakawa (2007). He compare the bias properties to the three types of GMM estimators: first differences, level and system. With that analysis, he concludes that the bias of the system GMM estimator is a weighted sum of the bias of the first two (GMM level and first difference). The bias is directed in the opposite direction. Therefore, the bias of first differences and level partly cancel each other out and the system GMM will closer to being consistent.

We consider a follow-up study conducted using the optimal weighting matrix: the GMM estimator under the assumptions of system instrumental variables (SIV.1-SIV.4)[13] is more efficient within the class of instrumental variable estimators. The standard approach in the literature on NIM is to simply not correct for this inconsistency. For example, Valverde and Fernández (2007) assume homoskedastic error terms, which could underestimated the standard error. Moreover, they consider the Sargan test, although the difference of Sangan tests is called for. We specify the pure Blundell and Bond (1998) to exploit the initial conditions as moment conditions (see also Blundell and Bond (2000)). The model we propose is specified as a simplified version of Eq.(7) :

$$y_{it} = \rho y_{i,t-1} + B_{it}^{'}\delta + u_{it},\ i = 1, ..., N;\ t = 2, \ldots, T. \tag{8}$$

Thinking somewhat ahead, a potential problem with the conventional procedure of linear system GMM concern the problem of too many instruments. Even when GMM is judged to be the appropriate estimation technique, we may question the following: are our instruments is valid instruments?

To answer this question, define $B_{it} \equiv (X_{it}, Z_t)$. Consider the simplified model Eq.(8). For the FE estimator we have: $E(\mu_i) = E(v_{it}) = E(\mu_i v_{it}) = 0$. Then, $\Delta y_{it} = \rho\Delta y_{i,t-1} + \Delta B_{it}^{'}\delta + \Delta v_{it}$ is instrumented by $y_{it} = \rho y_{i,t-1} + B_{it}^{'}\delta + u_{it}$ and $y_{it} = \rho y_{i,t-1} + B_{it}\delta + u_{it}$ is instrumented by $\Delta y_{it} = \rho\Delta y_{i,t-1} + \Delta B_{it}^{'}\delta + \Delta v_{it}$. This is the idea behind system GMM. The two equations are estimated simultaneously. The interesting thing here is the fact that a set of internal instrument used. For example $y_{i,t-2}$ instruments $\Delta y_{i,t-1}$ and $\Delta y_{i,t-1}$ instruments$y_{i,t-2}$. The resulting matrix of instrument is a stack of

[13]See Wooldridge (2010, Ch.8) for more discussion and notation used.

block-matrices:

$$w_o^t = \begin{bmatrix} w_o^1 & 0 & 0 & 0 & \cdots \\ 0 & w_0^2 & 0 & 0 & \cdots \\ 0 & 0 & w_0^3 & 0 & \cdots \\ 0 & 0 & 0 & w_0^4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \; and \; \triangle w_0^t = \begin{bmatrix} \triangle y_{i2} & 0 & \vdots & \vdots \\ 0 & \triangle y_{i3} & 0 & 0 \\ 0 & 0 & \triangle y_{i4} & 0 \\ \vdots & \vdots & 0 & \triangle y_{i5} \\ 0 & 0 & 0 & \vdots \end{bmatrix} \quad for\, t \geqslant 2.$$

In case of $t = 3$, $w_o^3 = (y_{i3}, y_{i2}, y_{i1})$, a difference GMM equation generates only one instrument per variable and the system GMM produces two instruments. If T increases, the instrument can quickly grow large relatively to the sample size, making asymptotic results about parameter and test statistic misleading. Specifically, there are at least two problems: firstly, is the feasible efficient GMM, in which sample moment are used to estimate an optimal weighting matrix, see Wooldridge (2010), to identifying moments between instruments and error. The sizes and dimension of the tests can be invalid. Secondly, IV without any restriction, can overestimated endogenous variables and biasing vector of parameter estimated towards those from non-instrumenting estimator.

Various authors analyss these problems, e.g. Arellano (2003), and show that the size is overestimated for non-endogenous variables at the order $O(k/N)$, k being the instrument count. Roodman (2009a) shows empirically that initial conditions of mean stationarity is not necessary.

The problem of proliferation of instruments can be overcome by limiting the number of lags, so as to not use all $W \equiv (w_0, \triangle w_0)$ available. Secondly, the system GMM instruments are collapsed into a stacked vector.

Below we show the results of running GMM on the T period of data. To keeping the comparability of the result, in both case the lags are collapse including Two Stages Least Square (2SLS) and robust. Nevertheless, the number of instruments perform equal and Hansen test of overidentication of restriction perform between value where we can qualify the parameter estimated unbiased and consistency. Country dummies are included, but time-dummies are not for reasons discussed earlier.

The results of the estimations are shown in the Table(5.4.1). The lagged variable on the right-hand side is positive and significant, thereby demonstrating the importance of accounting for previous values of the dependent variable. In accordance with Blundell and Bond (2000) the result corroborated, $\rho = 1$, the system GMM estimator gains improvements in the precision and also reduce bias in the finite sample. The estimates show that increases in NIM_{t-1} in one percent, ceteris paribus, NIM_{t-1} increases by 0.77530%.

In the case of comtemponeous effects, some covariates are revealed to have a significant effect: cost-to-income (cti_t), Equity over Total Assets ratio ($Etar_t$), off balance sheet item (Oba_t) and interest rate risk ($irrisk_t$). Strangely, management quality/efficiency affects NIM negatively effect and significantly. It is expected that a higher $Etar_t$ will reduce the banks need for external resources, and therefore lead to higher NIM. Finally, the $irrisk_t$ effect is found to significantly poisitive, it also contributes to augment the deposits rate spreads.

TABLE 5.4.1. Dynamic Model

Dependent Variable NIM	
Variables	Parameter
L.nim	0.77530***
cost	-0.00379*
loan	-0.00149
lloss	0.00012
etar	0.05289***
teatr	-0.01341
oba	0.00003***
crisk	-0.00082
irrisk	0.00090**
liqrisk	-0.00212
gdpg	-0.27890
hipc	3.82935
Constant	0.07113
Observations	731
Number of nr	196
Wald test	1626
Wald test p-value	0
Number Instruments	51
Hansen J	36.59
Hansen p-value	0.0480
Dif Hansen J	13.1
p value	0.362
AR(2) test	-0.286
AR(2) P-value	0.775

Robust Standard *** p<0.01, ** p<0.05, * p<0.1
Fixed Effect included

CHAPTER 6

Dealing With Attrition

As mentioned in the introduction, not all variables are observed for all banks. Moreover, not all banks are observed in all periods, due to attrition. We will try to deal with this attrition below. Note that attrition problems are not commonly is considered in the literature of banks net interest margin.

6.1. Inverse Probability Weighted (IPW) Estimation

We now consider the case study where the population of the biggest banks are restricted by the condition of "banks-alive". By, considering the banks-specific variables, this implies that only the dependent variable becomes fully observed. Even under this condition, it does not mean that there is a problem of attrition caused by the vector not fully observed, this is x_{it}.

There are a number of procedures to deal with attrition that use overly strong assumptions. One of those is known under the name "listwise deletion". This method drops all observations that contain missing values, and is also known as complete case analysis (CC). This procedure leads to inconsistent parameter estimates, unless the missing data are Missing Completely At Random [1]. A slightly weaker assumption is missing at random (MAR)[2].

The Litle and Rubin (2002) mechanism is useful to classify the missing data pattern, however it is a testable assumption and, could leads to correction procedures in the dataset, assuming that the observables are MAR. This procedure provides a good accessible treatment with an easy application, when data is missing only on one vector of the variable, for example, y_{it}, x_{it}, z_t, only vector x_{it} contains missing (not fully observed). As shown in Table(A2) the percentage of missing values varies over countries between maximum 39.75% and minimum 14.71% whithout any restriction (banks-alive).

[1]The data on x_{it} is said to be MCAR if the probability of missing data on x_{it} depends neither on its own values nor on the values of other variables in the dataset.

[2]For any vector x_{it} $(i = 1, ..., N,\ t = 1, ..., T)$ we call the missing at random, if values not observed in x_{it} does not depend on its value but may depend on the values of $x_{it}(t \neq s)$.

The problem of attrition is presents in the covariates, x_{it}. Various methods where proposed by several authors to deal with this issue and they depend on the structure of the missing pattern. However, many of these methods are handled under MAR. With slight similarity in some studies, the reader can be referred to Robins et al. (1994) for the details regarding the method for various missing data pattern. And for more complicated cases, see the simulation analysis for partial linear model with missing covariates Qin et al. (2012)[3]. Notice that, even though different equational models are used regarding the estimation technique, it is necessary to weight the estimating parameters.

In the presence of selection bias, we need to provide which conditions might be used to estimate the parameters consistently as in the form of the specification of the Chapter(5.4), and it also a special case of partial linear model.

For the attrition correction mentioned above, we can determine in which context a linear model generalized method of moment (GMM) estimator can be used for an estimation strategy, both belonging to the M-estimation and compatible with IPW. It is important to highlight that GMM is a particular case of the M-estimation and it is important to keep all processes of estimation in the same context, so there is consistency in terms of assumptions, see Hayashi (2000, p.206). However, we should consider all conditions of the process of estimation given in the Chapter (5.4). Do these problem persists when there is missing data?

In fact, the instrument is only an additionally problem to the GMM estimator. First, we need to analyze the moment conditions only for the observations on units for which a complete time series is available (subsample subjected to the attrition correction). Second, it is well known that estimators that base themselves on either the balanced subpanel(observables) or the fully unbalanced panel without correcting for selectivity bias, may result in biased estimates if the missing data is MAR.

Nonetheless, taking into consideration the raw of variables which provide us with two possibilities to choose the selection predictors y_{it} or z_t, both variables are fully observed. In this case, amongst others, sometimes the selection of a predictor cannot be fully observed. Motivated by efficiency, we highlight below the methodology in which the selection probabilities depends on predictors y_{it} or z_t are completely observed in

[3]These analysis concern of the nonlinear cases, in fact, when $g(T_i) = 0$ the model becomes a linear particular case. See also, Wang (2009)

all periods of time ($t = 1, \ldots, T$). Thus, this allow us to generate more efficient estimators for the estimation of the unknow probabilities of attrition.

Let $w_{it} \equiv (y_{it}, x_{it})'$ be a random vector of $M \times 1$ analyzing variables and taking values in $W \subset R^M$. Let $w_{it},\ i = 1, ..., N$ be a random vector of the same dimension. We assume that the population distribution F is that for some sequence drawn from unknowns distribution of population F. Let α be a $K \times 1$ unknown parameter vector, and let $F(w_{it}, \alpha)$ be a known vector-valued function of $\alpha \in \Delta \subset R^K$.

Assumption 6.1: i) The whole of population of the largest banks is represented by F.

ii) and F^1 $(y_{it}|x_{it}, s_{it} = 1)$, and for identification we need $E[F^1(w_{it}u_{it})] = 0$.

Assumption 6.2: The vector $w_{it} \equiv (y_{it}, x_{it})'$ is observed whenever $s_{it} = 1$, and we allows the possibility that y_{it} is observed along w_{it}.

Assumption 6.3: We assume the vector w_{it} results are random, meaning that the pattern missing data mechanisms do not depend of vector x_{it} and y_{it}. In addition, choosing all vector z_t as a selection indicator, z_t is observed whenever $s_{it} = 1, P(s_{it} = 1|w_{it}, z_t) = P(s_{it} = 1|z_t) \equiv p(z_t)$.

In the most important assumption, 6.1 part (ii), we define F^1 as the entire subpanel in which all individuals observed are included and consider the fully observed dependent variable. This allows for missings only in the x_{it}. The second consideration is assumption is 6.3, nominally Litle and Rubin (2002) denoted as MAR, which under that assumption allow us to estimate by Inverse Probability Weighting (IPW). Additionally, the estimation of weights involves certain regularity conditions such that moments of population exist and are finite, as well as continuity and differentiability of the log-likelihood of the linear function of F^1. Note that only this objective function will be appear in the minimization of the problem, since the observability condition for the dynamic model, $s^*_{it} \equiv 1[s_{it} \times s_{i,t-1} \times s_{i,t-2} = 1]$, allows all possible combination of the block matrix observed account for the estimation, guaranteeing that the selection of observables does not reduce too much the number of observations. That being said, we assume that $z_t, (y_{it}, x_{it})$ is independent of s^*_{it} from assumption 6.3, so that ignorability becomes

$$(s^*_{it} = 1|y_{it}, x_{it}, z_t) = P(s^*_{it} = 1|z_t). \tag{9}$$

The objective function[4] is parametricly defined by $F^1(w_{it}, \alpha_1)$, where the distribution of w depends on $P \times 1$ parameter space $\mathrm{A} \subset R^P$. The objective function has a unique solution for the population minimization; that solution may be $\alpha_1 \in \mathrm{A}$. This assumption implies that F^1 is correctly specified and that our distribution satisfies the moment condition. In this case, α_1 solves both the issues of misspecification and the problem of identification. Afterwards, we can proceed in solving the linear problem as follows:

$$min_{\alpha_1 \in A} \sum_{i=1}^{N} F^1(w_{it}, \alpha_1), \tag{10}$$

where the $F^1(.,.)$ contains the vector from Eq.(3) which intends to $E(x_{it}u_{it}) \neq 0$ and $E(z_t u_{it}) = 0$. In some cases it is necessary to specify correctly the conditional mean $E(y_{it}|x_{it}, s^*_{it}) = E(y_{it}|x_{it})$ for consistency estimation of α_1, but in this case it is not necessary given the random sample, since $s^*_{it}.F(w, \alpha) \equiv F^1(w_i, \alpha_1)$ is random. Thus, we cannot skip the rank condition, $E(x'_{it}x_{it}) = K$. Since $E(x'_{it}x_{it})$ is symmetric $K \times K$ matrix, which is the same assumption as $E(x'_{it}x_{it})$ being positive definite. Additionally, we need uniform convergence in probability, where the expected value of the estimator α_1 converge sample average $N^{-1}\sum_{i=1}^{N}[F^1(w_i, \alpha_1)] = \overline{F}^1(\alpha_1)$. If α_1 is obtained from a random sample it is different of true parameter $\alpha_1 \neq \alpha$, meaning α_1 is a candidate of estimator for α.

The previous discussion suggests some general consideration of M-Estimation (ME)[5], and it is applied when covariates are unobserved. Under some regularity condition ME estimator is consitent and asymptotic normal, as in special case as it applied in F^1.

Even under the missing at random (MAR)[6] assumption, because of the of dimensionality of matrix of instruments, the behaviour of estimators of β in finite samples becomes biased unless one imposes additional restrictions on either the missing completely at random (MCAR)[7] mechanism on the non/or semiparametric model for F^1. Hence the best that can be hoped for is an estimator that is GMM under the

[4]Note that F^1 is subsample of our population, so we skip the theory about whole population, see Wooldridge (2002).

[5]See Wooldridge (2010, ch.12).

[6]For any vector x_{it} $(i = 1, ..., N, t = 1, ..., T)$ we call the missing at random, if values not observed in x_{it} does not depend on its value but may depend on the values of $x_{it}(t \neq s)$.

[7]The data on x_{it} is said to be MCAR if the probability of missing data on x_{it} depends neither on its own values nor on the values of other variables in the data set.

MAR assumption can be biased, even when empirical moments condition based in $N^{-1}\sum_{i=1}^{N}[F^1(w_i,\alpha_1)] = \overline{F}^1(\alpha_1)$ is close to zero. In fact, the definition of s_{it}^* also induces poor instruments. Assuming the population moment holds,

$$E[F(w_{it},\alpha)] = 0, \tag{11}$$

and additionaly under ignorability condition, the GMM estimator in Eq.(11) result in selected sample analogues as in the population moments of the form,

$$E[s_{it}^* F(w_{it},\alpha)] = 0. \tag{12}$$

Using the notation that α_0 from the probit model, the weighted moment function in Eq.(12) by the inverse of the selection probability does hold if the expected value of the weighted selected population moment is equal as in Eq.(11), then

$$E\left[\frac{s_{it}^*}{P(z_t,\alpha_0)}F(w_{it},\alpha)\right] = 0. \tag{13}$$

In this setting, Wooldridge (2010, ch.12 and 14) established consistency and asymptotic normality of the limit distribuition of the GMM. However, the context is relatively different, since it does not envolve weights as in Eq.(13). GMM is sufficiently general to accomodate a large class of function in econometrics, it is questionable whether it is relevant to apply IPW under ignorabilibity, and one needs to assure that the moment condition in Eq.(11) translated into the Eq.(13) holds. More precisely, the value of α_1 solves Eq.(13), where we saw, $s_{it}^* F(w_{it},\alpha) = F^1$. Using all the statements underlying Eq.(9), are evident and clear in answering to the question. This can be seen as follows:

- $E[s_{it}^* F(w_{it},\alpha)] = E\{E[F^1(w_{it},\alpha_1)|z_t]\}$, using Law of Iterated Expectations
  $= E\{E(s_{it}^*|z_t)E[F^1(w_{it},\alpha_1)|z_t]\}$, under ignorability
  $= E\{P(z_t,\alpha_0)E[F^1(w_{it},\alpha_1)|z_t]\}$, using the result $E(s_{it}^*|z_t) = P(z_t,\alpha_0)$, and $s_{it}^* F(w_{it},\alpha) = F^1(.,.)$, then if we weighting unless multiply by scalar, it becomes
- $E\left[\frac{s_{it}^*}{P(z_t,\alpha_0)}F(w_{it},\alpha)\right]$, where is a scalar $P(z_t,\alpha_0) > 0$ for $t = 1,...,T$.
  $= E\left\{E\left[\frac{s_{it}^*}{P(z_t,\alpha_0)}F(w_{it},\alpha)|z_t\right]\right\}$
  $= E\left[\frac{1}{P(z_t,\alpha_0)}E(s_{it}^*|z_t)E\{F(w_{it},\alpha)|z_t\}\right] = E\left[\frac{1}{P(z_t,\alpha_0)}P(z_t,\alpha_0)E\{F(w_{it},\alpha)|z_t\}\right]$
  $= E\left[E\{F(w_{it},\alpha)|z_t\}\right]$
  $= E\left[F(w_{it},\alpha)|z_t\right] = 0.$

It is now clear that the moment conditions of the selected sample may not hold. However, the weighted selected sample population moment conditions do hold. These two points differ in terms of set of moment condition and using the weighted moment condition is equivalent as in Wooldridge (2010, ch.12 and 14). Again, it is intuitively clear that under two moment conditions it is necessary to hold MCAR in the first point and MAR in the second. This leads to our decision to choose the weighted selected sample, point two.

The two-step dynamic NIM equation was estimated; in the first stage we run a probit regression s_{it}^* on collected covariates of country-specific, say z_t, to estimate the the probabilities. In the Second stage, the estimated IPW is included in the dynamic model of NIM to obtain the parameter using the GMM estimator. This meaning, the resolution of attrition problem, initiate with a setting the balanced panel(unweighted) and then applying the IPW to obtain weighted estimator.

The estimation results are displayed in Table(6.1.1). To summarize, these results are the Inverse Probability Weighted GMM estimation results. They exploit the moment conditions of the population model, that are weighted to correct for possible selection due to attrition. For comparability with the estimaton results obtained earlier, in both cases the lags are collapsed (including 2SLS) and robust. The number of instruments is the same and the Hansen test of overidentication of restrictions does not reject the model. The interpretation of the model estimates, this time with accounting for attrition, is the same as before. The estimate of the coefficient of oba_t=00002% is positive and significant but can hardly be considered substantive in the economic sense.

TABLE 6.1.1. Dynamic Model with IPW Correction

Dependent Variable NIM		
Variables	MCAR	MAR-IPW
L.nim	0.86326***	0.84266***
cost	-0.00814***	-0.01154***
loan	0.00462	0.00987**
lloss	0.00231	0.00419
etar	0.04866***	0.03822***
teatr	-0.01016	-0.00805
oba	0.00001	0.00002*
crisk	0.00502	0.01183**
irrisk	0.00076**	0.00081**
liqrisk	-0.00402*	-0.00080
gdpg	-0.33635	-0.12075**
hipc	4.41162**	3.55710*
Constant	-0.19559	-0.40634
Observations	426	426
Number of nr	143	143
Wald test	1374	1221
Wald test p-value	0	0
Number Instruments	51	51
Hansen J	25.70	19.30
Hansen p-value	0.368	0.736
Dif Hansen J	12.7	8.54
p value	0.392	0.742
AR(2) test	0.0356	0.0370
AR(2) P-value	0.972	0.971

Robust Standard *** p<0.01, ** p<0.05, * p<0.1

Fixed Effect included

CHAPTER 7

Conclusion

This paper has studied the influence of bank-specific and macroeconomic variables on European bank interest margins. A range of model specifications have been discussed. The static linear regression model and linear panel data model with random effects to be inadequate. The latter model was rejected even under relatively weak exogeneity assumptions under the null hypothesis. The standard fixed effects model was first generalized to allow for endogeneity due to correlation of regressors with the idiosyncratic disturbance term u_{it}. A dynamic version of this model was then estimated. The last extension amounted to Inverse Probability Weighting of the GMM moment conditions, to correct for potential inconsistencies due to non-random attrition.

- The GMM estimation based on the IPW has the best chance of generating consistent estimates.
- We find that when the dynamic panel data model is not corrected for attrition, some estimates have signs that are contrary what one would expect.
- The macroeconomic effect of GDP growth has a negative impact on the net interest margin of banks, and hipc has a positive effect. Both effects are statistically significant.

TABLE 7.0.1. GMM vs GMM_{IPW}

Variables	CC	MCAR	MAR-IPW
L.nim	0.77530***	0.86326***	0.84266***
cost	-0.00379*	-0.00814***	-0.01154***
loan	**-0.00149**	0.00462	0.00987**
lloss	0.00012	0.00231	0.00419
etar	0.05289***	0.04866***	0.03822***
teatr	-0.01341	-0.01016	-0.00805
oba	0.00003***	0.00001	0.00002*
crisk	**-0.00082**	0.00502	0.01183**
irrisk	0.00090**	0.00076**	0.00081**
liqrisk	-0.00212	-0.00402*	-0.00080
gdpg	**-0.27890**	-0.33635	-0.12075**
hipc	**3.82935**	4.41162**	3.55710*
Constant	**0.07113**	-0.19559	-0.40634

The CC model is the estimates from the Table(5.4.1).

The MCAR estimates is the s^*_{it} monotonic restriction for banks alive .

The MAR-IPW estimates is the s^*_{it} monotonic restriction for banks alive .

Bibliography

Ahn, S. C. & Low, S., 1996. A reformulation of the hausman test for regression models with pooled cross-section-time-series data. Journal of Econometrics 71 (1-2), 309 – 319.

Allen, L., 1988. The determinants of bank interest margins: A note. Journal of Financial and Quantitative Analysis 23, 231–235.

Arellano, M., 2003. Modeling optimal instrumental variables for dynamic panel data model. Centro de Estudios Monetarios y Financieros, Madrid (0310).

Baltagi, B. H., 2005. Econometric Analysis of Panel Data. New York: Wiley.

Baltagi, B. H., Li, Q., 1992. A note on the estimation of simultaneous equations with error components. Econometric Theory 8 (1), 113–119.

Baltagi, B. H., Liu, L., 2007. Alternative ways of obtaining hausman's test using artificial regressions. Statistics & Probability Letters 77 (13), 1413 – 1417.

Beck, T., Demirgüç-Kunt, A., Levine, R., 2006. Bank concentration, competition, and crises: First results. Journal of Banking and Finance 30, 1581–1603.

Blundell, R., Bond, S., 1998. Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics 87, 115 –143.

Blundell, R., Bond, S., 2000. Gmm estimation with persistent panel data: An application to production functions. Econometric Reviews 19, 321–340.

Claeys, S., Vennet, R. V., 2008. Determinats of bank interest margins in central and eastern europe: a comparasion with west. Economic systems 32 (2), 197–216.

Demirgüç-Kunt, A., Laeven, L., Levine, R., 2003. Regulations, market structure, institutions, and the cost of financial intermediation. Journal of Money, Credit and Banking 36 (3), 593 – 622.

Hahn, J., Ham, J., Moon, H. R., 2011. Test of random effect versus fixed effects with small with variation. Economics Letters 112, 293 –297.

Hausman, J., 1978. Specification test in econometrics. Econometrica 46 (6), 1273 –1291.

Hayakawa, K., 2007. Small sample bias properties of the system {GMM} estimator in dynamic panel data models. Economics Letters 95 (1), 32 – 38.

Hayashi, F., 2000. Econometrics. Princeton University Press.

Ho, T. S. Y., Saunders, A., 1981. The determinants of bank interest margins: Theory and empirical evidence. Journal of Financial and Quantitative Analysis 16 (4), 581–600.

John Bound, D. A. J., Baker, R. M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. Journal of the American Statistical Association 90, 443–450.

Kasman, A., Tunc, G., Vardar, G., Okan, B., 2010. Consolidation and commercial bank net interest margins: Evidence from the old european union members and candidate countries. Economic Modeling 27, 648 –655.

Litle, R. J. A., Rubin, D. B., 2002. Statistical Analysis with Missing Data. New York: Wiley.

Maudos, J., de Guevara, J. F., 2004. Factors explaining the interest margin in the banking sectors of the european union. Journal of Banking and Finance 28, 2259–2281.

Mundlak, Y., 1978. On the Pooling of Time Series and Cross Section Data. Econometrica 46 (1), 69–85.

Nelson, C. R., Startz, R., 1990b. Some further results on the exact small sample properties of the instrumental variables estimator. Econometrica 58, 967–976.

Nickel, S. J., 1981. Biases in dynamic models with fixed effects. Econometrica 49, 1417–1426.

Phillips, P. C., Sul, D., 2007. Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. Journal of Econometrics 137 (1), 162 –188.

Qin, G., Zhu, Z., Fung, W. K., 2012. Robust estimation of the generalised partial linear model with missing covariates. Journal of Nonparametric Statistics 24 (2), 517–530.

Robins, J. M., Rotnitzky, A., Zhao, L. P., 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89 (427), 846–866.

Roodman, D., 2009a. A note on the theme of too many instruments*. Oxford Bulletin of Economics and Statistics 71 (1), 135–158.

Saunders, A., Schumacher, L., 2000. The determinants of bank interest rate margin: An international study. Journal of international Money and Finance 19 (6), 813 – 832.

Staiger, D., Stock, J. H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586.

Valverde, S. C., Fernández, F. R., 2007. The determinants of bank margins in european banking. Journal of Banking and Finance 31 (7), 2043–2063.

Wang, Q.-H., 2009. Statistical estimation in partial linear models with covariate data missing at random. Annals of the Institute of Statistical Mathematics 61 (1), 47–84.

Wooldridge, J. M., 2002. Inverse probability weighted {M}-estimators for sample selection, attrition, and stratification. Portuguese Economic Journal 1 (2), 117–139.

Wooldridge, J. M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT Press.

Appendix

TABLE A1. Variables names

Type of Variables	Label	Description
Banks specific	NIM	[(Interest Income-Interest Expense)/ (Average Interest Bearing Assets)]*100
	Cost	(Operating cost/(Interest Income-Expense + Other operating Income))*100
	Loan	(Net Loan/ Total Asset)*100
	Lloss	Loan Loss Provision/Gross Loan
	Etar	Equity/ Total Asset
	Teatr	Total Earning Asset/ Total Assets
	Oba	Off Balance Item/Total Earning asset
	Crisk	Loan default/Total Loan
	Irrisk	Interbank market interest rate (Tree months)
	Liqrisk	Liquid Asset/Short Term Found
country specific	Tot	Capacity to import less exports goods
	M2	Average annual growth rate in money and quasi money.
	Hipc	Is calculated in accordance with harmonized statistical methods
	Unem	Percentage of Labor force
	Gdpg	Annual percentage growth rate of GDP at market price constant
	Inves	Total investment Percent of GDP
	Savg	Gross national savings Percent of GDP
	Expen	General government total expenditure Percent of GDP
	Cac	Current account balance Percent of GDP

TABLE A2. Summarize and Misstable

Variable	a=Obs=.	b=Obs<.	Total(a+b)	%Obs=.	values Unique	Mean	Std,Dev,	Min	Max
nim	275	1594	1869	14.71%	>500	1.56%	1.17%	-1.233%	15.43%
cost	287	1582	1869	15.34%	>500	59.94%	31.53%	.574%	547.89%
loan	282	1587	1869	15.08%	>500	57.73%	23.04%	0.00%	99.81%
lloss	513	1356	1869	31.19%	>500	5.24%	14.34%	-320.75%	194.57%
etar	278	1591	1869	14.87%	>500	5.90%	7.45%	-30.52%	100.00%
teatr	308	1561	1869	14.48%	>500	79.83%	34.42%	.037%	100.00%
oba	743	1126	1869	39.75%	>500	51.89%	508.44%	0.00%	13847.93%
crisk	476	1393	1869	25.47%	>500	1.28%	5.29%	0.00%	813.65%
irrisk	439	1430	1869	23.49%	>500	105.20%	130.17%	0.00%	943.43%
liqrisk	356	1513	1869	19.05%	>500	49.88%	86.73%	.129%	974.99%

TABLE A3. Number Observation by country

country	N° Banks	nim	cost	loan	loan	lloss	etar	teatr	oba	crisk	irrisk	liqrisk
AT	22	132	133	133	133	104	133	125	57	114	107	115
BE	13	84	84	84	84	77	84	79	43	81	70	84
DE	34	214	209	214	214	171	214	214	185	189	193	194
DK	14	92	92	92	92	63	92	92	85	52	82	84
ES	13	75	75	75	75	61	75	61	47	60	61	61
FI	13	72	72	73	73	49	73	73	61	47	62	66
FR	28	161	162	155	155	149	162	162	101	154	155	156
GB	26	158	157	158	158	119	158	155	130	131	133	153
GR	13	88	88	88	88	88	88	88	53	88	81	88
IE	13	78	76	77	77	65	78	78	60	71	63	78
IT	26	142	142	141	141	128	143	143	83	140	137	143
LU	13	78	77	78	78	73	78	78	68	78	78	78
NL	13	61	58	61	61	53	55	55	26	52	50	55
PT	13	75	74	75	75	74	75	75	53	79	78	79
SE	13	84	83	83	83	82	83	83	74	57	80	79
Total	267	1594	1582	1587	1587	1356	1591	1561	1126	1393	1430	1513

TABLE A4. Mean values of the variables

country	nim	cost	loan	loan	lloss	etar	teatr	oba	crisk	irrisk	liqrisk
AT	1.965	58.141	57.512	57.512	0.670	11.182	94.293	14.489	0.690	105.202	79.883
BE	1.607	68.842	49.091	49.091	0.257	5.518	95.144	17.293	6.834	128.761	26.095
DE	1.128	64.444	55.885	55.885	33.812	3.281	97.404	7.827	0.684	82.269	42.304
DK	1.130	43.042	70.905	70.905	1.320	5.424	95.566	6.939	0.478	90.487	93.324
ES	1.851	50.822	68.961	68.961	0.699	6.064	93.083	25.573	0.609	105.848	19.265
FI	1.165	69.639	60.960	60.960	0.320	8.941	93.288	42.239	2.345	127.846	40.173
FR	1.597	70.004	42.696	42.696	2.199	4.305	0.899	0.162	0.874	117.758	63.997
GB	1.445	60.591	50.435	50.435	0.763	4.529	92.291	27.726	0.969	111.617	68.498
GR	3.026	64.954	65.612	65.612	1.251	7.687	91.304	21.331	1.055	137.400	24.153
IE	0.854	44.457	57.258	57.258	1.556	4.000	98.151	12.370	0.484	100.896	73.323
IT	2.152	58.995	66.199	66.199	2.824	7.185	93.650	20.857	0.614	95.732	36.305
LU	0.908	46.306	28.178	28.178	0.316	4.871	94.615	601.876	3.205	102.423	49.012
NL	1.072	70.024	59.142	59.142	0.314	4.033	95.001	13.996	0.695	76.839	37.484
PT	1.858	68.193	66.349	66.349	0.551	5.934	92.947	47.258	0.801	128.807	30.567
SE	1.529	48.917	82.606	82.606	0.002	7.982	0.977	0.367	0.358	86.398	30.567

TABLE A5. Standard Deviation values of the variables

country	nim	cost	loan	loan	lloss	etar	teatr	oba	crisk	irrisk	liqrisk
AT	1.836	29.126	25.485	25.485	0.624	16.324	4.703	9.642	0.316	155.716	179.082
BE	0.610	25.732	22.103	22.103	0.571	7.392	6.156	21.217	19.339	146.994	24.285
DE	0.886	39.535	18.973	18.973	25.978	1.707	2.492	7.176	0.469	100.590	48.060
DK	0.577	21.288	17.912	17.912	1.736	2.741	4.406	7.907	0.317	140.513	155.285
ES	0.465	7.729	8.305	8.305	0.397	1.418	2.106	7.531	0.132	111.632	11.158
FI	0.622	53.557	31.164	31.164	0.712	14.966	7.961	231.098	7.171	158.368	51.553
FR	1.374	26.473	20.664	20.664	1.035	2.244	0.178	0.101	0.725	119.770	42.251
GB	1.105	27.621	23.345	23.345	1.112	2.183	8.361	42.295	0.831	108.073	106.878
GR	0.717	17.895	12.759	12.759	1.203	7.742	3.182	19.714	0.541	187.777	40.349
IE	0.606	32.227	24.986	24.986	5.391	4.239	2.107	14.656	0.328	144.604	154.896
IT	0.818	27.756	14.517	14.517	1.873	2.830	3.650	21.517	0.280	106.716	33.305
LU	0.473	18.741	16.968	16.968	1.217	2.825	7.719	1988.907	6.263	64.107	20.618
NL	0.384	38.178	16.815	16.815	0.402	1.900	3.319	8.964	0.520	63.397	39.160
PT	0.740	35.890	16.395	16.395	1.224	8.860	3.986	55.920	0.726	174.333	25.808
SE	1.977	26.410	13.897	13.897	0.006	8.969	0.019	1.067	0.385	143.738	25.808